Analysis of Boyer and Moore's MJRTY algorithm

Laurent Alonso, Edward M. Reingold

Analysis of Boyer and Moore's `MJRTY` Algorithm

Laurent Alonso* Edward M. Reingold†

April 11, 2013

Abstract: Given a set of n elements each of which is either red or blue, Boyer and Moore's `MJRTY` algorithm uses pairwise equal/not equal color comparisons to determine the majority color. We analyze the average behavior of their algorithm, proving that if all 2^n possible inputs are equally likely, the average number of color comparisons used is $n - \sqrt{2n/\pi} + O(1)$ with variance $(\pi - 2)n/\pi - \sqrt{2n/\pi} + O(1)$.

Key words: Algorithm analysis, majority problem

Given a set $\{x_1, x_2, \ldots, x_n\}$, each element of which is colored either red or blue, we must determine an element of the majority color by making equal/not equal color comparisons $x_u : x_v$; when n is even, we must report that there is no majority if there are equal numbers of each color. How many such questions are necessary and sufficient? It is easy to obtain an algorithm using at most $n-\nu(n)$ questions, where $\nu(n)$ denotes the number of 1-bits in the binary representation of n; furthermore, $n-\nu(n)$ is a lower bound on the number of questions needed (see [**?**] and [**?**]). In [**?**], the average case was investigated: Assuming all 2^n distinct colorings of the n elements are equally probable, $\frac{2n}{3} - \sqrt{\frac{8n}{9\pi}} + O(\log n)$ comparisons are necessary and sufficient in the average case to determine the majority.

In this note we analyze the average performance of an inferior, but historically important majority algorithm, that of Boyer and Moore [**?**], first described in 1980 as an example of the logic supported by the Boyer-Moore Theorem Prover (1971; see [**?**]). The algorithm, Algorithm 1, is subtle in that it works correctly even when there are arbitarily many colors *as long as the set is known to have an element that occurs more than half the time* (in this multi-color case, if the set is not known to have a majority element, a second pass over all n

*INRIA-Lorraine and LORIA, Université Henri Poincaré-Nancy I, BP 239, 54506 Vandoeuvre-lès-Nancy, France. Email: `Laurent.Alonso@loria.fr`

†Department of Computer Science, Illinois Institute of Technology, 10 West 31st Street, Chicago, Illinois 60616-2987, USA. Email: `reingold@iit.edu`

elements is necessary). It is Algorithm 1, called MJRTY in [?], that led to an intensive study of the matter; see [?, sec. 5.8] and [?] for historical details.

Algorithm 1 Boyer and Moore's MJRTY algorithm [?].

```
1:  c ← 0
2:  for i ← 1 to n do
3:     // if c = 0 there are equal numbers of red/blue items among
4:     // x_1, ..., x_{i-1}; otherwise there are c more of the color of x_j
5:     if c = 0 then
6:        j ← i
7:        c ← 1
8:     else if x_i = x_j then
9:        c ← c + 1
10:    else
11:       c ← c − 1
12:    end if
13: end for
14: if c = 0 then
15:    No majority
16: else
17:    Majority is color of x_j
18: end if
```

In the two-color case, the number of color comparisons in line 8 of MJRTY is equal to n less the number of times the algorithm finds $c = 0$ in line 5, so we focus on counting that number of times. But $c = 0$ at line 5 can happen only when i is odd, which happens at least once (when $i = 1$) and at most $\lceil n/2 \rceil$ times. It happens when $i = 2k+1$ if $x_1, \ldots, x_{2k}$ contains k copies of each of the two colors, which happens with probability $\binom{2k}{k}/2^{2k}$ if all 2^n two-colorings of $x_1, \ldots, x_n$ are equally probable. Let $m = \lceil n/2 \rceil$; the average number of times that $c = 0$ in line 5 is thus

$$\sum_{k=0}^{m-1} \binom{2k}{k} / 2^{2k}, \tag{1}$$

partial sums of the central binomial coefficients, divided by corresponding powers of 4. The generating function for the central binomial coefficients is $C(z) = 1/\sqrt{1-4z}$, so the generating function for the partial sums in (1) is $C(z/4)/(1-z) = 2C'(z/4)$; hence

$$\sum_{k=0}^{m-1} \binom{2k}{k} / 2^{2k} = m\binom{2m}{m} / 2^{2m-1}, \tag{2}$$

where $m\binom{2m}{m}$ are called Apéry numbers [?, A005430]. By Stirling's formula, the righthand side of (2) is $\sqrt{2n/\pi} + O(1)$ for $m = \lceil n/2 \rceil$. It follows that in the two-color case Boyer and Moore's MJRTY algorithm uses at least $\lfloor n/2 \rfloor$ and at

most $n-1$ color comparisons, and an average of

$$n - \sqrt{2n/\pi} - O(1) \tag{3}$$

color comparisons.

We now compute the variance. Let $S(n)$ be the set of all possible 2^n two-color input sequences to `MJRTY`. For $I \in S(n)$ and $1 \le i \le n$, let

$$c_i(I) = \begin{cases} 1 & \text{if } c = 0 \text{ in line 5 at iteration } i \text{ for input } I, \\ 0 & \text{otherwise,} \end{cases}$$

and let $S_i(n) \subseteq S(n)$ be the set of input sequences I such that $c = 0$ at iteration i for input I; and hence $c_i(I) = 1$ for all $I \in S_i(n)$. With this notation, our derivation of the average (3) is

$$\begin{aligned} \sum_{I \in S(n)} \left(n - \sum_{i=1}^{n} c_i(I) \right) \mathbf{Pr}(I) &= n - \sum_{I \in S(n)} \left(\sum_{i=1}^{n} c_i(I) \right) \mathbf{Pr}(I) \\ &= n - \sqrt{2n/\pi} - O(1). \end{aligned} \tag{4}$$

The variance is

$$\begin{aligned} &\sum_{I \in S(n)} \left(n - \sum_{i=1}^{n} c_i(I) \right)^2 \mathbf{Pr}(I) - \left(\sum_{I \in S(n)} \left(n - \sum_{i=1}^{n} c_i(I) \right) \mathbf{Pr}(I) \right)^2 \\ &\quad = \sum_{I \in S(n)} \left(\sum_{i=1}^{n} c_i(I) \right)^2 \mathbf{Pr}(I) - \left(\sum_{I \in S(n)} \left(\sum_{i=1}^{n} c_i(I) \right) \mathbf{Pr}(I) \right)^2, \end{aligned} \tag{5}$$

because $\mathbf{Var}(x - y) = \mathbf{Var}(x) + \mathbf{Var}(y) - 2\mathbf{CoVar}(x, y)$ and we have $x = n$ which is constant. Thus we need the value of

$$\sum_{I \in S(n)} \left(\sum_{i=1}^{n} c_i(I) \right)^2 \mathbf{Pr}(I) = 2^{-n} \sum_{I \in S(n)} \left(\sum_{i=1}^{n} c_i(I) \right) \left(\sum_{j=1}^{n} c_j(I) \right).$$

which, because $\mathbf{Pr}(I) = 1/|S(n)| = 2^{-n}$, and by distributivity,

$$= 2^{-n} \sum_{I \in S(n)} \sum_{i=1}^{n} c_i(I) \left(\sum_{j=1}^{n} c_j(I) \right),$$

so changing the order of summation,

$$\begin{aligned} &= 2^{-n} \sum_{i=1}^{n} \sum_{I \in S(n)} c_i(I) \left(\sum_{j=1}^{n} c_j(I) \right), \\ &= 2^{-n} \sum_{i=1}^{n} \sum_{I \in S_i(n)} \left(\sum_{j=1}^{n} c_j(I) \right) \end{aligned}$$

because $c_i(I)$ is 1 if $I \in S_i(n)$ and 0 otherwise,

$$= 2^{-n} \sum_{i=1}^{n} ||S_i(n)||, \tag{6}$$

where $||S_i(n)||$ is the total number of times that $c = 0$ at line 5 over all inputs $I \in S_i(n)$.

Note that for $m \geq 1$, in the input sequence $x_1, \ldots, x_{2m}$, the two possible choices for the color of x_{2m} do not affect the number of times that $c = 0$ at line 5 (because $x_1, \ldots, x_{2m-1}$ must contain a majority color, hence $c \neq 0$ in line 5 for $i = 2m$). Thus we need only compute $||S_i(n)||$ for odd n because $||S_i(2m)|| = 2||S_i(2m-1)||$.

We can view an arbitrary odd-length input sequence $x_1, \ldots, x_n \in S_{2k+1}(n)$, $n = 2m - 1$, as being composed of two contiguous subsequences: a even-length front part $x_1, \ldots, x_{2k}$, in which there is no majority color, and an odd-length rear part $x_{2k+1}, \ldots, x_n$. These subsequences are independent, so we can count the number times $c = 0$ in each separately.

First, consider the possible front parts, sequences $x_1, \ldots, x_{2k}$ with no majority color. The total number of times that $c = 0$ at line 5 over all such sequences is

$$f_k = \sum_{j=0}^{k-1} \binom{2j}{j} \binom{2(k-j)}{k-j} = \sum_{j=0}^{k} \binom{2j}{j} \binom{2(k-j)}{k-j} - \binom{2k}{k}, \tag{7}$$

because if $c = 0$ happens at line 5 for $i = 2j + 1$ (with $j < k$), neither of the subsequences: $x_1, \ldots, x_{2j}$ and $x_{2j+1}, \ldots, x_{2k}$ (which is not empty) contains a majority color. Observing that the sum in (7) is a convolution of the central binomial numbers with themselves, the generating function for this sum is $C(z)^2 = 1/(1 - 4z)$ and hence

$$f_k = 4^k - \binom{2k}{k}.$$

Now consider the rear part, an arbitrary odd-length input sequence x_1, ..., x_{2k+1}. The total number of times that $c = 0$ in line 5 over all such sequences is

$$r_k = \sum_{j=0}^{k} \binom{2j}{j} 2^{2(k-j)+1} = 2^{2k+1} \sum_{j=0}^{k} \binom{2j}{j} / 2^{2j},$$

because $c = 0$ at line 5 for $i = 2j + 1$ if $x_1, \ldots, x_{2j}$ contains j elements of each color and the remaining $2(k - j) + 1$ elements are colored arbitrarily,

$$= (k+1)\binom{2k+2}{k+1},$$

evaluated as in (2).

Finally, because each possible front word occurs $2^{2m-1-2k}$ times as first part of a word in $S_{2k+1}(2m-1)$, and each rear word occurs $\binom{2k}{k}$ times, we obtain

$$||S_{2k+1}(2m-1)|| = 2^{2m-1-2k} f_k + \binom{2k}{k} r_{m-1-k}.$$

Therefore,

$$||S_{2k+1}(2m-1)|| = 2^{2m-1} - 2^{2m-1}\binom{2k}{k}/2^{2k} + \binom{2k}{k}\left((m-k)\binom{2(m-k)}{m-k}\right),$$

and so we have

$$\begin{aligned} 2^{-2m+1}\sum_{i=1}^{2m-1} ||S_i(2m-1)|| &= 2^{-2m+1}\sum_{k=0}^{m-1} ||S_{2k+1}(2m-1)|| \\ &= \sum_{k=0}^{m-1} 1 - \sum_{k=0}^{m-1}\binom{2k}{k}/2^{2k} \\ &\quad +2^{-2m+1}\sum_{k=0}^{n-1}\binom{2k}{k}\left((m-k)\binom{2m-2k}{m-k}\right). \end{aligned}$$

We can simplify the second sum as in (2); for the third sum, we note that it is a convolution,

$$C(z)(zC'(z)) = \frac{2z}{(1-4z)^2} = \frac{z}{2}\frac{d}{dz}\left(\frac{1}{1-4z}\right) = \sum_{j=1}^{\infty} j 2^{2j-2} z^j,$$

so that

$$\begin{aligned} 2^{-2m+1}\sum_{i=1}^{2m-1} ||S_i(2m-1)|| &= 2^{-2m}\sum_{i=1}^{2m} ||S_i(2m)|| \\ &= 2m\left(1 - \frac{\binom{2m}{m}}{4^m}\right). \end{aligned} \tag{8}$$

We have $m = \lceil n/2 \rceil$, so putting together (2), (5), (6), and (8), Stirling's formula gives the variance of the number of color comparisons in line 8 of `MJRTY`,

$$2m\left(1 - \binom{2m}{m}/4^m\right) - \left(m\binom{2m}{m}/2^{2m-1}\right)^2 = \frac{\pi-2}{\pi}n - \sqrt{2n/\pi} + O(1).$$